DOCUMENT RESUME

ED 204 415 TM 810 499

AUTHOR Wilcox, Rand R.
TITLE Determining Test Length to Control for False-Positive and False-Negative Error Rates on Criterion-Referenced Test.
INSTITUTION California Univ., Los Angeles. Center for the Study of Evaluation.
REPORT NO CSE-R-147
PUB DATE Oct 80
NOTE 50p.

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Bayesian Statistics: Criterion Referenced Tests: *Latent Trait Theory: *Mathematical Models: *Scores: True Scores
IDENTIFIERS Binomial Error Model: Test Length

ABSTRACT

Concern about passing those examinees who should pass, and retaining those who need remedial work, is one problem related to criterion-referenced testing. This paper deals with one aspect of that problem. When determining how many items to include on a criterion-referenced test, practitioners must resolve various non-statistical issues before a particular solution can be applied. A fundamental problem is deciding which of three true scores should be used. The first is based on the probability that an examinee is correct on a "typical" test item. The second is the probability of having acquired a typical skill among a domain of skills, and the third is based on latent trait models. Once a particular true score is settled upon, there are several perspectives that might be used to determine test length. These solutions are reviewed and critiqued. Some new results are described that apply when latent structure models are used to estimate an examinee's true score. (Author/GK)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED204415

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
[X] This document has been reproduced as received from the person or organization originating it.
[ ] Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

DETERMINING TEST LENGTH TO CONTROL FOR FALSE-POSITIVE AND FALSE-NEGATIVE ERROR RATES ON CRITERION-REFERENCED TESTS

Rand R. Wilcox

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

_____

_____

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

CSE Report No. 147

October, 1980

Test Design Project
Center for the Study of Evaluation
Graduate School of Education, UCLA
Los Angeles, California 90024

TM 810 499

Table of Contents

	Page
Preface	i
Abstract	ii
1. Introduction	1
2. The Purpose of the Test	2
3. Solutions Using Domain Scores, π_0 Known	3
A Bayesian Approach	9
An Alternative Approach	11
Error Rates for the Typical Examinee	14
4. Solutions Using Domain Scores, π_0 Unknown	17
Bayesian Solutions When π_0 is Unknown	22
5. Solutions in Terms of Proportion of Skills Acquired	23
6. Solutions Using Latent Structure Models	28
7. Solutions in Terms of ξ and a Population of Examinees	31
Bounds on the Probability of an Error	33
8. Solutions Using Latent Trait Models	34
A Concluding Remark	37
References	38

PREFACE

A part of our goal at CSE has been to develop new and improved psychometric techniques to study, develop and characterize achievement tests and achievement test items. Recently our efforts have been focused on certain errors that occur when using criterion-referenced tests. In particular, we have investigated problems related to estimating and controlling the false-positive and false-negative error rates associated with a test and a population of examinees. In other words, we are concerned about passing those examinees who should pass, and retaining those examinees who need remedial work. This paper deals with one aspect of that problem.

ABSTRACT

When determining how many items to include on a criterion-referenced test, practitioners must resolve various non-statistical issues before a particular solution can be applied. A fundamental problem is deciding which of three true scores should be used. The first is based on the probability that an examinee is correct on a "typical" test item. The second is the probability of having acquired a typical skill among a domain of skills, and the third is based on latent trait models. Once a particular true score is settled upon, there are several perspectives that might be used to determine test length. The paper reviews and critiques these solutions. Some new results are described that apply when latent structure models are used to estimate an examinee's true score.

1. Introduction

When trying to determine how many items to include on a criterion-referenced test, perhaps the most fundamental problem is that there are at least three conceptualizations or models of an achievement test that might be used. Each of these conceptualizations is based on a different type of true score. The first deals with the number of items an examinee would get correct if he/she were to respond to every item in some item domain. The second is concerned with the proportion of skills among a domain of skills that an examinee has acquired. Because of errors at the item level, such as guessing, this conceptualization is different from the first. The final approach is based upon latent trait models. In some cases, one model might yield substantially different results from another in terms of test length, and so the choice of a model can be crucial.

Once one of the above conceptualizations is settled upon, a variety of other issues must be resolved. For example, when comparing an examinee's true score to a standard, do we assume the standard is known, or do we want to take into account the process by which it was determined? Do we formulate the test length problem in terms of a single examinee, a "typical" examinee, or both? How certain do we want to be of making a correct decision (classification) of an examinee? Are we willing to use a Bayesian solution?

The first goal of this paper is to give a brief review and critique of the three general approaches that might be used when determining the length of a criterion-referenced test. In addition, new results on test length are described. Finally, possible directions for future research are indicated.

2. The Purpose of the Test

Consistent with earlier test length solutions, it is assumed that the purpose of the test is to sort the examinees into one of two mutually exclusive groups. In addition, it is assumed that it is possible to define these two groups in terms of some notion of true score, say π, that characterizes a particular examinee. For the moment, the exact nature of the true score, π, need not be specified.

Let π_0 be a constant which may or may not be known. We refer to π_0 as the standard. An examinee is said to belong to the first group, which is designated as S_G, if his/her true score is greater than or equal to π_0. If $\pi<\pi_0$, the examinee is said to belong to S_B. The problem is to determine how many items to include on the test so that we can be reasonably certain of correctly determining whether an examinee belongs to S_G or S_B.

We should mention that a variety of real-life situations exist where it is desired to sort examinees into one of two mutually exclusive groups. In some cases we want to determine mastery of a specific skill or a narrowly defined set of objectives. In other instances, interest centers on proficiency in a variety of skills that characterize a particular subject area. The point is that the term criterion-referenced test might be deemed inappropriate for certain situations. For the present, this is a minor

issue. The important idea is that we want to compare π to the constant π_0. We should also mention that Hambleton et al. (1978) describes two primary uses of criterion-referenced tests, namely, estimating domain scores, and classifying examinees. Our only concern is with the latter use.

3. Solutions Using Domain Scores, π_0 Known

In the context of a mastery or criterion-referenced test, perhaps the most frequently used notion of true score is based on the concept of an item sampling model (e.g., Harris, 1974; Huynh, 1976; Novick and Lewis, 1974; Wilcox, 1977). Consider, for example, a domain of dichotomously scored items. In some cases, the item pool actually exists while, in other cases, the notion of an item domain is a convenient conceptualization. For this model, π represents the proportion of items an examinee would answer correctly if he/she were to answer every item in the item pool. For a single examinee responding to a random sample of n items, the probability of getting x correct is assumed to be

$$\binom{n}{x} \pi^x (1-\pi)^{n-x}, \qquad [3.1]$$

the binomial probability function. This is justified when items are randomly sampled from an infinite item pool, or a finite pool with replacement, and π remains constant. Using equation 3.1 is also appropriate when it gives a good fit to the observed scores of an examinee. Gelfand and Thomas (1976), Katz (1963), Tarone (1979) and Cochran (1954) discuss the problem of determing when a good fit is obtained.

One difficulty with this model occurs when we consider more than one examinee. If we let $g(\pi)$ represent the distribution of true scores over a population of examinees, equation 3.1 implies that the marginal distribution of observed scores is given by

$$\int_0^1 \binom{n}{x} \pi^x (1-\pi)^{n-x} g(\pi)\, d\pi. \qquad [3.2]$$

Lord and Novick (1968, section 23.8) show that equation 3.2 implies that the correlation between observed scores and true scores is given by the KR21 reliability formula. If every examinee takes the same n items, the implication is that every item has the same level of difficulty. This result prompted Lord and Novick to replace equation 3.1 with a two-term approximation to the more general compound binomial model, the approximation being given by

$$P_n(x) + d\pi\,(1-\pi)\,C(x) \qquad [3.3]$$

where

$$P_n(x) = \binom{n}{x} \pi^x (1-\pi)^{n-x} \qquad [3.4]$$

and

$$C(x) = \sum_{v=0}^{2} (-1)^{v+1} \binom{2}{v} P_{n-2}(x-v) \qquad [3.5]$$

The parameter d is equal to

$$\frac{n^2(n-1)\sigma_\rho^2}{2\{\mu_x(n-\mu_x) - \sigma_x^2 - n\sigma_\rho^2\}} \qquad [3.6]$$

where μ_x^2 and σ_x^2 are the mean and variance, respectively, of the marginal distribution of observed scores and where σ_ρ^2 is the variance of the item difficulties. It should be noted that if every examinee takes a different random sample of n items, the simpler binomial probability function is theoretically justified and equation 3.2 is correct. Further comments concerning equation 3.3 are made below.

Apparently, the earliest attempt at providing a solution to the test length problem was made by Millman (1973) using the binomial probability function. Shortly thereafter, Fhaner (1974) gave a more formal approach again using the binomial probability function but with an indifference zone built into the analysis. This means that in addition to the known constant π_0, a constant $\delta^* > 0$ is specified with the idea that if the examinee's percent correct true score is less than or equal to $\pi_0 - \delta^*$ or greater than or equal to $\pi_0 + \delta^*$, we want to be reasonably certain of making a correct decision. If $\pi_0 - \delta^* < \pi < \pi_0 + \delta^*$ any decision is said to be correct.

The goal can be stated more precisely as follows: Let n_0 be a specified passing score, i.e., if the examinee's observed score is greater than or equal to n_0, the decision $\pi \geq \pi_0$ is made; otherwise we decide that $\pi < \pi_0$. The problem is to determine the smallest n, so that regardless of the actual value of π, the probability of a correct decision (CD) is reasonably high, say greater than or equal to P^*. More briefly, we want

$$P(CD) \geq P^*, \ 1/2 < P^* < 1. \qquad [3.7]$$

The reason for requiring $P^* > 1/2$ is that we can guarantee that the P(CD) is at least .5 without any observations at all, simply by randomly deciding whether π is above or below π_0.

For $\pi<\pi_0$ we have that

$$P(CD) = \sum_{x=0}^{n_0-1} \binom{n}{x} \pi^x (1-\pi)^{n-x} \qquad [3.8]$$

and for $\pi\geq\pi_0$

$$P(CD) = \sum_{x=n_0}^{n} \binom{n}{x} \pi^x (1-\pi)^{n-x} \qquad [3.9]$$

Moreover, it can be shown that for $\pi\leq\pi_0-\delta^*$, equation 3.8 is minimized at $\pi=\pi_0-\delta^*$ for any n and that for $\pi\geq\pi_0+\delta^*$, equation 3.9 is minimized at $\pi=\pi_0+\delta^*$. Thus, to satisfy equation 3.7 for any π, it is sufficient to find the smallest n so that for $\pi=\pi_0-\delta^*$, equation 3.8 exceeds P^* and simultaneously for $\pi=\pi_0+\delta^*$, equation 3.9 also exceeds P^*. Note that for $\delta^*=0$, both equations 3.8 and 3.9 approach .5. Thus, $\delta^*>0$ is a necessary condition for ensuring that an n exists satisfying equation 3.7.

It is also of interest to observe that it is relatively easy to incorporate virtually any loss function into the above framework. Suppose, for example, $L_1(\pi)\geq 0$ is the loss associated with mis-classifying an examinee for whom $\pi<\pi_0$ and let $L_2(\pi)\geq 0$ be the loss when $\pi\geq\pi_0$. The risk or expected loss is

$$L_1(\pi) \sum_{x=x_0}^{n} \binom{n}{x} \pi^x (1-\pi)^{n-x}, \quad \pi<\pi_0 \qquad [3.10]$$

$$L_2(\pi) \sum_{x=0}^{n_0-1} \binom{n}{x} \pi^x (1-\pi)^{n-x}, \quad \pi\geq\pi_0 \qquad [3.11]$$

Using numerical procedures, the values of π maximizing equations 3.10 and 3.11

are readily determined. Moreover, if there is an open interval around π_0 such that $L_1(\pi)=L_2(\pi)=0$, and if L_1 and L_2 are bounded above, both equations 3.10 and 3.11 can be made arbitrarily small.

Before concluding this sub-section we note that Wilcox (1979a) has generalized Fhaner's solution in two directions. In particular, Wilcox's solution applies to any model involving some notion of true score π (not necessarily domain scores) for which there exists a statistic $\hat{\pi}(x)$ for estimating π such that the cumulative distribution function of $\hat{\pi}(x)$, say $F(\hat{\pi}(x)|\pi)$, is stochastically increasing. In other words, it is assumed that $\pi<\pi'$ implies that $F(\hat{\pi}(x)|\pi')\leq F(\hat{\pi}(x)|\pi)$ for all x. Consider any group of k examinees and let g be the number of examinees for whom $\pi\geq\pi_0$. It follows that over all possible configurations of true score, the minimum probability of a correct decision is given by

$$\prod_{i=1}^{k-g} F(\pi_0|\pi_i=\pi_0-\delta^*) \qquad \prod_{j=k-g+1}^{k} 1-F(\pi_0|\pi_j=\pi_0+\delta^*) \qquad [3.12]$$

where $\pi_1,\ldots,\pi_{k-g}$ are the true scores of the k-g examinees for whom $\pi<\pi_0$ and $\pi_{k-g+1},\ldots,\pi_k$ are the true scores for the examinees having $\pi\geq\pi_0$. It has been shown that in terms of g, equation 3.12 is minimized at g=0 or g=k if $\hat{\pi}_i(x)$ (the statistic for estimating π_i) is independent of $\hat{\pi}_j(x)$, $i\neq j$. Thus, by examining these two cases and choosing n accordingly we can guarantee equation 3.7 no matter what the values of the π_i's happen to be. Wilcox's solution contains the binomial error model, Poisson process models and normal distributions as a special case.

Finally, in the case of percent correct true score, Wilcox (1979a) indicates that the simpler binomial model appears to give a conservative solution when the conditional distribution of observed scores is given by a two-term approximation to the compound binomial model as described by equation 3.3. In other words, the binomial error model appears to result in a longer test length than would be obtained using equation 3.3, all other things be equal. However, a rigorous proof that this is the case has not been derived.

An important feature of the test length solution proposed by Fhanér (1974) and extended by Wilcox (1979a) is that it is conservative in the sense that it makes no assumption about the value of the examinee's true score π. Furthermore, it is assumed that there is no information beyond an examinee's observed score for making the decision about whether π is above or below π_0. For a single examinee, Fhanér's solution might result in the use of a moderate number of items on the test. Suppose, for example, we assume the binomial error model holds, we set k=1, P^*=.9, δ^*=.1, π_0=.8 and the passing score n_0 is chosen to be the smallest integer such that $n_0/n \geq .8$. It follows that n=26 is the shortest test length satisfying equation 3.7.

In practice, however, it is often desired to simultaneously make a decision about k>1 examinees. If we insist on using a conservative approach to the test length problem and if we view the decision making process in terms of k examinees, the minimum probability of a correct decision decreases rapidly as k gets large. (See Wilcox, 1979a.)

For k large, one might argue that it is conservative but unrealistic to consider the case where the value of every examinee's true score

is equal to $\pi_0-\delta^*$ or $\pi_0+\delta^*$. Thus, some other perspective might be deemed more appropriate when judging the adequacy of the length of the test. The remainder of this section considers how this might be done.

A Bayesian Approach

Novick and Lewis (1974) describe a Bayesian approach to determining n that applies to the case of a single examinee whose conditional distribution of observed scores is given by the binomial probability function. As is typically done for the binomial case, an examinee's true score is viewed as a random variable with a distribution belonging to the beta family. More specifically, it is assumed that the probability density function of π is given by

$$\frac{\Gamma(r+s)}{\Gamma(r)\Gamma(s)} \pi^{r-1}(1-\pi)^{s-1} \qquad [3.13]$$

where r>0 and s>0 are unknown parameters and Γ is the usual gamma function. If r and s were known, it might be possible to justify a shorter test length than would be required if the approach used by Fhanér (1974) were employed. Another appealing feature of the Bayesian solution is that we would know the probability of a correct decision given the examinee's observed score.

More specifically, the test length solution is formulated as follows: Given equation 3.13, and assuming equation 3.1 holds, it is known that the conditional distribution of π given an observed score x, is

$$h(\pi|x) = \frac{\Gamma(n)\pi^x(1-\pi)^{n-x}}{\Gamma(x+1)\Gamma(n-x+1)} \qquad [3.14]$$

This is a beta distribution with parameters x+1 and n-x+1. Thus, we can compute the probability of $\pi \geq \pi_0$ when $x=n_0$. The test length is determined by increasing n until this probability is believed to be reasonably close to one.

Novick and Lewis also illustrate how to incorporate a simple loss function into their analysis. In particular, let a be the "loss" of passing an examinee who should fail, and let b be the cost of failing an examinee who should pass. We advance a student if $bP(\pi \geq \pi_0 | x, n) \geq aP(\pi < \pi_0 | x, n)$. Note that we need only specify the ratio a/b, not the actual values of a and b, since we are comparing the ratio a/b to $Pr(\pi \geq \pi_0 | x, n)/Pr\pi < \pi_0 | x, n)$.

We observe that Morgan (1979) has extended the work of Novick and Lewis (1974) to situations where guessing and carelessness are incorporated into the analysis. Novick (1973) discusses the specification of the prior distribution. Novick and Lewis (1974) state that the specification of the prior must be done carefully. They go on to suggest that the book by Novick and Jackson (1974) and the paper by Novick, Lewis and Jackson (1973) might aid in this process.

Perhaps most issues in statistics are controversial, at least to some degree. Consider, for example, the problem of estimating the mean of a distribution. The sample mean has various optimal properties under certain circumstances (e.g., normality) but a variety of alternative estimates might be used instead (Andrews, et al., 1972). When discussing Bayesian solutions to a problem, it seems prudent to remind the reader that this area of statistics is a bit more controversial than others. M. S. Bartlett, commenting on a paper by D. V. Lindley in a book edited by Godambe and Sprott (1971, p. 447), writes as follows:

> "I would say that the statisticians' model is different in principle from a prior distribution in that it can be tested. Where it cannot be tested this is to me unsatisfactory. Prior distributions are, as I understand it, in general untestable. What does Professor Lindley mean when he says that 'the proof of the pudding is in the eating'? If he has done the cooking it is not surprising if he finds the pudding palatable, but what is his reply if we say that we do not. If the Bayesian allows some general investigation to check the frequency of errors committed, or even real losses, this might be set up; but if the criterion is inner consistency, then to me this is not acceptable."

Despite Bartlett's comment, the importance of Bayesian statistics should not be underestimated. Even if one insists on the classical approach, Bayesian methods may prove to be valuable (e.g., Murray 1977). For further favorable comments toward the Bayesian approach to statistical inference, the reader is referred to Kendall and Stuart (1973, pp. 159-161).

An Alternative Approach

There is an approach to statistical inference developed by Dempster (1966, 1967) which might be applied to the test length problem. Apparently this approach has not been discussed in terms of the problem at hand and so, for completeness, we describe it here.

Dempster's results are quite general but, for the sake of clarity, the discussion is limited to the case of a single examinee for whom the binomial error model applies.

Suppose $\pi_0=.7$, $n=10$, $n_0=7$ and that an examinee's observed score is $x=6$. Thus, the decision $\pi<\pi_0$ would have been made. If the observed score had been $x=1$, say, the same decision would have been made but one might "feel" more certain the correct decision had been reached. Assuming that we should feel more certain about the decision when $x=1$ versus $x=6$, the question arises as to how to express this certainty in some meaningful way. For the Bayesian statistician, the problem is relatively straightforward since, once the beta prior has been specified, $P(\pi<\pi_0|x=1)$ and $P(\pi<\pi_0|x=6)$ can be calculated.

Dempster's theory does not give us an exact value for $P(\pi<\pi_0|x)$; rather it yields two values, say P_1 and P_2, which are interpreted as lower and upper bounds, respectively, on $P(\pi\leq\pi_0|x)$. Bounds on $P(\pi\geq\pi_0|x)$ can also be derived.

From Dempster (1968a) it can be seen that for $P(0\leq\pi\leq\pi_2|x)$,

$$P_2 = \sum_{y=x}^{n} \binom{n}{y} \pi_0^y \, (1-\pi_0)^{n-y} \qquad [3.14]$$

and

$$P_1 = \sum_{y=x}^{n-1} \binom{n}{y} \pi_0^y \, (1-\pi_0)^{n-y} \; . \qquad [3.15]$$

As for $P(\pi_0 \leq \pi \leq 1 | x)$

$$P_2 = \sum_{y=0}^{x} \binom{n}{y} \pi_0^y (1-\pi_0)^{n-y} \qquad [3.16]$$

$$P_1 = \sum_{y=0}^{x-1} \binom{n}{y} \pi_0^y (1-\pi_0)^{n-y}. \qquad [3.17]$$

In terms of test length, one might choose an n so that P_1 is reasonably close to one for $P(0 \leq \pi \leq \pi_0 | x < x_0)$ and $P(\pi_0 \leq \pi \leq 1 | x \geq x_0)$. If, however, an examinee's observed score is x_0, it can be seen that such an n may not exist. One way out of this dilemma is to incorporate an indifference zone into the analysis but in a slightly different fashion than was done in the earlier portion of this paper. Here, we might look at the bounds on $P(0 \leq \pi \leq \pi_0 + \delta^* | x)$ and $P(\pi_0 - \delta^* \leq \pi \leq 1 | x)$. In this case, we would be choosing an n so that

$$\sum_{y=n_0-1}^{n-1} \binom{n}{y} (\pi_0 + \delta^*)^y (1-\pi_0-\delta^*)^{n-y}, \qquad [3.18]$$

the lower bound on $P(0 \leq \pi \leq \pi_0 - \delta^* | x = x_0 - 1)$,

and

$$\sum_{y=0}^{n_0-1} \binom{n}{y} (\pi_0 - \delta^*)^y (1-\pi_0+\delta^*)^{n-y}, \qquad [3.19]$$

the lower bound on $P(\pi_0 - \delta^* \leq \pi \leq 1 | x = n_0)$, are reasonably close to one.

In practice it would seem that this approach might yield nearly the same results as those obtained with the classical methods described earlier. However, one might prefer Demspter's approach to the usual Bayesian solution because it is possible to incorporate into the analysis prior beliefs about whether π is above or below π_0 without specifying a specific form for the prior. Dempster (1968a) illustrates

how this might be done. For further comments on this approach to making inferences, the reader is referred to the discussion following the paper by Dempster (1968b).

Error Rates for the Typical Examinee

In addition to considering the adequacy of the test length in terms of a single examinee or k>1 specific examinees, one might consider the examinees being tested as a random sample from some larger population of examinees and consider whether n is sufficiently large for the typical examinee being tested (Wilcox, 1977). (See, also, Huynh, 1980; Livingston, 1979.) It is not being suggested that the analysis presented in this subsection replace the approaches described above. Rather, the results reviewed and outlined here might be used to give us additional insight into whether there are an adequate number of items on the test.

We now assume that observed test scores x_i (i=1,...,k) are available for k examinees and we consider the estimation of $\alpha=P(x\geq x_0,\ \pi<\pi_0)$, the probability of a false-positive decision, and $\beta=P(x<x_0,\ \pi\geq\pi_0)$, the probability of a false-negative decision for a randomly selected examinee. Since we have observations for estimating α and β, the present approach might be termed a "retrospective" study. This is in contrast to Fhanér (1974) and Wilcox (1979a) where it is assumed that no information is available concerning an examinee's true score and so the problem is more along the lines of designing an experiment.

Note that π is again an unknown fixed constant for a specific examinee; no prior distribution is considered for an examinee as is done in the Bayesian solution. We refer to a distribution for π say $w(\pi)$,

but now the distribution of π is over a population of examinees.

Let $h(x|\pi)$ be the conditional distribution of observed scores for an examinee having true score π. It follows that

$$\alpha = \sum_{x=n_0}^{n} \int_0^{\pi_0} h(x|\pi)w(\pi)d\pi \qquad [3.20]$$

and

$$\beta = \sum_{x=0}^{n_0-1} \int_{\pi_0}^{1} h(x|\pi)w(\pi)d\pi. \qquad [3.21]$$

If we want to define α and β in terms of an indifference zone, we simply replace π_0 with $\pi_0-\delta^*$ in equation 3.20 and we replace π_0 with $\pi_0+\delta^*$ in equation 3.21.

If we knew $h(x|\pi)$ and $w(\pi)$ we would know α and β. If α and β were judged to be too large, we could decrease their values by increasing the test length.

For most situations, neither $h(x|\pi)$ nor $w(\pi)$ is known. Suppose, however, we follow Lord and Novick (1968, Chapter 23) and assume $h(x|\pi)$ is some approximation to the compound binomial distribution. Further suppose that the moments of the true score distribution can be estimated using the x_i's. Lord (1965) describes how to do this when the two-term approximation to the compound binomial given by equation 3.3 is deemed to be appropriate. Once these estimates are available, the methods described below for estimating $w(\pi)$ can be employed.

Perhaps the most frequently used approach to estimating the true score distribution is to assume that $w(\pi)$ belongs to the family of beta distributions (e.g., Keats and Lord, 1963; Lord, 1965). If the true

score distribution is unimodal, a good approximation to it may be possible with a beta distribution (Springer, 1979, p. 268). If the true score distribution is multimodal, it is not clear when-or even if-a good fit to the true score distribution can be obtained. One possible problem is that (excluding U-shaped distributions) beta distributions can have, at most, one mode. To complicate matters, there is no satisfactory method of detecting a poor fit to the true score distribution using the observed x_i's. The difficulty is that even if we are given the first m moments of π over the population of examinees (m being any integer), the true score distribution is not uniquely determined. There are alternative approaches to estimating the distribution of π (e.g., Lord, 1969; Blischke, 1964; Maritz, 1970; von Mises, 1964, pp. 384-401) but the circumstances under which these procedures give more accurate results appears to be unknown.

Because of the difficulty in determining the accuracy of point estimates of α and β, there is some doubt as to when we should rely on such estimates when judging the adequacy of the test length. An alternative approach that might be used is to estimate bounds on α and β which make no assumptions about the shape of the true score distribution. Wilcox (1979c) indicates how this can be done. The solution is based on estimating the first two moments of π and applying results by Skibinsky (1977) which yield bounds on the probability that π is in a particular interval. It should also be pointed out that an earlier paper by Skibinsky (1976) describes how to use the first three moments of the true score distribution to obtain upper bounds to the probability that π is in the interval $(0, \pi_0)$ or $(\pi_0, 1)$. Following Wilcox (1979c)

these bounds might also be used to determine bounds on α and β.

We conclude this section by noting that the beta-binomial model appears to give a good estimate of α and β even when the conditional distribution of observed scores are generated according to the two-term approximation of the compound binomial given by equation 3.3 (Wilcox, 1977). Thus, when investigating the adequacy of the test length, the beta-binomial model would seem to suffice when the true score distribution belongs to the beta family. Moreover, a moderate number of examinees usually gives a reasonably accurate estimate of α and β when the beta-binomial model holds. However, there are occasions when observed scores on a moderate number of examinees can result in wildly inaccurate estimates of the parameters of the beta distribution (Wilcox, 1979f). This is a highly unusual event, but it seems prudent to keep this fact in mind when considering test length.

4. Solutions Using Domain Scores, π_0 Unknown

So far we have given a very brief outline and review of procedures for judging test length, all of which assume that the criterion score π_0, is known. In reality the criterion score is not known; rather, it is determined by some process. Huynh (1976), for example, describes a method for determining π_0 when an external criterion exists.

One important aspect of a criterion-referenced test is the effect the process of determining π_0 has on the test length, n. In other words, it may be of interest to incorporate this process into the analysis. This is done by Wilcox (1979b) for the case where π_0 is the unknown

parameter of some distribution. The examples given are based on the notion that the control (the distribution characterized by π_0) is a population of examinees. In this section, we consider a variation of this situation.

In practice, the criterion score is often specified by a panel of judges. In an attempt to better approximate reality, the following conceptualization is used. A total of k judges have specified a criterion score, π_{0i} (i=1,..., k). Furthermore, these k judges are viewed as a random sample from some population of individuals who are qualified for specifying π_0. In particular, it is assumed that the realization of π_{0i} is independent of π_{0j}, $i \neq j$ and that π_0 is the mean of the criterion scores that would be specified by the population of judges. Since π_0 is unknown we estimate it with $\bar{\pi} = k^{-1} \sum_{i=1}^{k} \pi_{0i}$. Accordingly, if an examinee's true score is estimated to be greater than or equal to $\bar{\pi}_0$ the decision $\pi \geq \pi_0$ is made; otherwise the reverse is said to be true.

Let $G(\bar{\pi})$ be the cumulative distribution function of $\bar{\pi}$ and consider the case of a single examinee for whom the binomial error model holds. It follows that the probability of a correct decision is given by

$$\int_0^1 \sum_{x=[n\bar{\pi}_0]}^{n} \binom{n}{x} \pi^x (1-\pi)^{n-x} \, dG(\bar{\pi}), \quad \text{if } \pi \geq \pi_0 \qquad [4.1]$$

or by

$$\int_0^1 \sum_{x=0}^{[n\bar{\pi}-1]} \binom{n}{x} \pi^x (1-\pi)^{n-x} \, dG(\bar{\pi}), \quad \text{if } \pi < \pi_0 \qquad [4.2]$$

where $[\bar{n}]$ is the smallest integer greater than or equal to $\bar{n}$.

Again it is neccssary to specify an indifference zone (i. e., a $\delta^* > 0$) to be certain that there exists an n so that equation 3.7 is satisfied. From Wilcox (1979b) it follows that the minimum probability of a correct decision is given by

$$\int_0^1 \sum_{x=0}^{[n\bar{\pi}-1]} \binom{n}{x} (\pi_0-\delta^*)^x (1-\pi_0+\delta^*)^{n-x} \, dG(\bar{\pi}) \qquad [4.3]$$

or

$$\int_0^1 \sum_{x=[n\bar{\pi}]}^{n} \binom{n}{x} (\pi_0+\delta^*)^x (1-\pi_0-\delta^*)^{n-x} \, dG(\bar{\pi}) \qquad [4.4]$$

whichever is smallest.

There remains the technical problem that $G(\bar{\pi})$ is unknown. One approach would be to assume that $G(\bar{\pi})$ is the distribution that minimizes the P(CD) so that no matter what the distribution of $G(\bar{\pi})$ happens to be, we can choose n so that $P(CD) \geq P^*$. A method of deriving this distribution is unknown to the author. However, comments made by Wilcox (1979b) suggest that if we assume that

$$P(\pi_{0i} = 0) = P(\pi_{0i} = 1) = 1/2 \qquad [4.5]$$

a reasonably conservative solution to the test length problem will be obtained. Note that this distribution is the limiting form of a non-informative beta prior used in Bayesian statistics. (See, e.g., Aitchison and Dunsmore, 1975, Chapter 2.)

There are three reasons for suspecting that the distribution given by equation 4.5 will give a conservative solution when specifying n. The first is that this distribution has the maximum possible variance of any

distribution on the closed interval [0, 1]. The second reason stems from considering the asymptotic case. Finally, familiarity with variational methods (e.g., Rustagi, 1976) suggests that the minimum of equations 4.3 and 4.4 over all possible distributions $G(\bar{\pi})$ occurs when $G(\bar{\pi})$ is a step function.

Note that when equation 4.5 holds, $G(\bar{\pi})$ is a binomial distribution. Thus, the approximate solution for specifying n that is given by Wilcox (1979b, expression (7)) can be applied to the present situation. Suppose, for example, $P^*=.9$ and $\delta^*=.1$. From Wilcox (1979b, Table 1), n=k=84 is required. In other words, if we administer n=84 items to an examinee and if we have k=84 judges specify a criterion score, there is (approximately) at least a 90% chance of correctly classifying the examinee.

In practice there might be at least two objections to the procedure just given. The first is that it might be too conservative in the sense that it is unrealistic to expect (or perhaps even allow) a judge to specify $\pi_0 = 0$ or 1. If we assume that every judge will specify a π_0 that is between .5 or .9, perhaps a fewer number of judges would be required. The second objection (related to the first) is that the variance of π_0 over judges might be small, relative to the variance of the observed score of an examinee, so that there is no need for sampling as many judges as items as was done for convenience in the illustration given above.

When comparing a single examinee's percent correct true score π to π_0 we may view our goal as determining which of two populations has the larger mean (i.e., we are trying to determine whether π is larger

or smaller than π_0). Thus, in the asymptotic case, we may apply the results given by Bechhofer (1954). We illustrate how this might be done.

For any random variable y having mean μ and variance σ^2 that is defined on the closed interval [a, b], $\sigma^2 \leq (\mu-a)(b-\mu)$ with equality holding when $P(y=a) = (b-\mu)/(b-a)$ and $P(y=b) = 1-P(y=a)$ (e.g., Skibinsky, 1977). Suppose, for the sake of illustration, we assume (or require) that $.5 \leq \pi_0 \leq .9$. It follows that the maximum possible variance of π_0 is .04 which occurs when $\pi_0 = .7$ and $P(\pi_0=.5) = P(\pi_0=.9) = .5$. Thus, in an attempt to find a conservative choice for n and k, we consider the case in which $\pi_0 = .7$ and the variance of π_0 is .04.

Suppose $P^* = .9$ and $\delta^* = .1$. Via the central limit theorem, we may apply the solution proposed by Bechhofer (1954, p. 24). In particular, the required number of judges is $k=d(.04)/(\delta^*)^2$ where d is read from Bechhofer's Table 1 (the column headed, in Bechhofer's notation, with k=2 and t=1). For $P^* = .9$, d=1.8124 and so, after rounding, k=13. As for n, firstly we observe that in our example, $\pi_0 - \delta^* = .6$ and $\pi_0 + \delta^* = .8$ (i.e., under the assumptions made, the P(CD) is minimized either when the examinee's true score is .6 or .8). Since a binomial distribution with probability of .6 has a larger variance than when $\pi=.8$, we consider the case $\pi=.6$. Thus, for this special case, the variance of a binomial distribution for a single observation is .24 and so $n \doteq (1.8124)^2(.24)/.01 \doteq 78$. This result also follows from Bechoffer's equation (34). (For related comments on the actual P(CD) in the case of normal distributions, see Tong and Wetzell, 1979; Lam and Chiu, 1976.)

As a partial check on the accuracy of the approximate solution for k and n, we used Monte Carlo procedures to estimate the P(CD) with k=13, n=78, $P(\pi_{0i}=.5)=P(\pi_{0i}=.9)=.5$ and $\pi=.6$. The resulting estimate was .912. As a further check, we used the approximate solution for $P^*=.75$ and .95. The corresponding values of (k, n) were (4, 22) and (22, 130), respectively. The estimated P(CD)'s were .76 and .94.

As a final comment, we note that when π_0 is known, the above illustrations indicate that the required number of items on the test is reduced considerably. For instance, suppose we know that $\pi_0=.7$. In this case, for a given P^* and δ^*, the minimum required test length is approximately equal to

$$\lambda^2.7(.3)/(\delta^*)^2 \qquad [4.6]$$

where λ is the P^* quantile of the standard normal distribution (Wilcox, 1979a). Thus, the values of n corresponding to $P^*=.75$, .9, .95 are approximately 10, 34 and 57 respectively. We see, therefore, that having precise information regarding π_0 can have a substantial effect on the test length.

Bayesian Solutions When π_0 is Unknown

It is possible to transform a binomial distribution to a normal distribution having known variance (e.g., Freeman and Tukey, 1950). If the distribution of π_0 is assumed to be normal with known variance, and if we apply the Freeman-Tukey transformation to the observed score of the examinee, the Bayesian approach described by Huang (1975) might be applied. However, the main results reported by Huang are concerned with finding optimal decision rules (Bayes procedures) for determining whether π is greater than, or less

than, π_0; no discussion is given on finding the smallest n so that equation 3.4 is attained.

5. Solutions in Terms of Proportion of Skills Acquired

In this section, we assume it is meaningful to say that an examinee either "knows" or "does not know" the answer to a particular item on a test. Alternatively, we might say that an examinee has, or has not, acquired the skill that is represented by a particular test item. Still another description of the approach taken here would be to say that the probability of a correct response to an item is a function of a dichotomized latent trait (Harris and Pearlman, 1978).

It should be stressed that when we describe an examinee as either knowing or not knowing the correct response to an item, no implication is being made that learning is all or none. Consider any model, for example, a latent trait model (see, e.g., Hambleton, et al., 1978) or classical test theory, in which the probability of a correct response is a function of some continuous unobservable variable. Either this variable has a value at which the examinee has a tendency to get the item right (the probability is greater than or equal to .5) or the examinee has a tendency to get it wrong. In some cases we might want to make inferences about this tendency as is the case in the Lazarsfeld-Kendall "turnover" model as described by Goodman and Kruskal (1959). No insistence is being made that such a continuous unobservable variable exists. The point is that describing an examinee as knowing or not knowing does not rule out, or have implications about some other continuous latent trait variable or model since we can always go from any latent trait to a latent state model. We should note, however, that the latent class point of view used in this section of the paper is deterministic in the sense that if we knew an examinees latent state, and <u>if</u> there were no errors at the item level, we could predict the

examinee's observed response. For this special case, we have, from the point of view of latent trait theory, Guttman item characteristic curves (cf. van der Linden, 1979). Reulecke (1977) and the references cited therein discuss, and further clarify, the relative merits of using latent classes in mental test theory, and so further comments are omitted.

There are two different but highly related approaches that have been considered, based on the framework just described. The first, which seems to have received the most attention in the literature, is to consider a specific skill in terms of a population of examinees (e.g., Harris and Pearlman, 1978; Marks and Noll, 1967). Macready and Dayton (1977) illustrate how this point of view can be used, among other things, to determine the number of items to be used when making a mastery/nonmastery decision concerning a particular skill. Their solution was recently extended by Bergan et al. (1980).

In this section, we concentrate on the second point of view which considers a single examinee in terms of a domain of skills. Let ξ be the proportion of skills that the examinee knows. Consistent with previous sections, the goal is to determine whether ξ is above or below some known criterion score ξ_0. The problem is to find a minimum value for n, the test length, so that regardless of the actual value of ξ, we are reasonably certain of making a correct decision whenever $\xi \leq \xi_0 - \delta^*$ or $\xi \geq \xi_0 + \delta^*$. If ξ is in the open interval $(\xi_0 - \delta^*, \xi_0 + \delta^*)$, i.e., the indifference zone, any decision is said to be correct.

One reason for considering this conceptualization of testing is that it occurs in real-life situations. For example, certain state-wide testing programs designed to determine a student's eligibility for graduation from high school have taken this view. A second reason is that it provides an interesting perspective on the test length problem. As alluded to earlier, formulating the problem in terms of ξ rather than the domain score π, can have a dramatic effect on the value of n. Finally, when measuring achievement

it seems reasonable to formulate the problem in terms ξ, the proportion of skills an examinee has acquired.

We consider two errors at the item level. They are

$$\gamma = P(\text{incorrect} \mid \text{examinee knows}) \qquad [5.1]$$

and

$$\varepsilon = P(\text{correct} \mid \text{examinee does not know}). \qquad [5.2]$$

From a frequentist point of view, we interpret ε as the proportion of correct responses an examinee would get among all the items in the item pool he/she does not know. Alternatively, one might define ε as the probability of a correct response to the same item over independent trials. This is similar to using the propensity distribution in classical test theory (Lord and Novick, 1968, Chapter 2) except that here, the distribution is defined in terms of an item an examinee does not know. To avoid the estimation problems noted by Wilcox (1979e), the former definition of ε is used. Of course, γ can be defined in an analogous fashion.

A fundamental problem with this approach to testing is deciding whether additional errors at the item level should be included in the analysis. Duncan (1974), for example, argues that in some cases, a misinformation model should be used. That is, we allow for the possibility that an examinee chooses an incorrect response to a multiple-choice test item because he/she believes it is, indeed, correct. Here, however, only the errors represented by equations 5.1 and 5.2 are considered.

Wilcox (1979d) has given some consideration to the relationship between γ, ξ and n, the test length. It was found that if one item per skill was used, an extremely large number of items might be needed to satisfy equation 3.7. Suppose, for example, $P^* = .9$, $\delta^* = .1$ and $\xi_0 = .8$. Further

suppose we are willing to assume that $.1 \leq \xi \leq .3$ and $0 \leq \gamma \leq .1$. In this case over 2600 items would be required to guarantee that $P(CD) \geq P^*$. The problem is that by allowing γ and ε to have positive values, we are shrinking the indifference zone in terms of the domain score π. One approach to this problem, which is considered by Wilcox (1979d), is to use more than one item per skill. It was found that this might lower the overall number of items on the test; however, a large number of items might still be required. We consider some alternative solutions.

In the case of multiple-choice test items, one possible approach is to use the usual correction for guessing formula score. (For a Bayesian formula score, see Molenaar, 1977.) Assuming one item per skill is used, and that each item has m alternatives from which to choose, the formula score is $x-(n-x)/(m-1)$ where, as before, x is the observed (number correct) score. This suggests we estimate ξ with

$$\hat{\xi} = n^{-1}[x-(n-x)/(m-1)] \qquad [5.3]$$

(See, also, van den Brink and Koele, 1980.)

Note that $\hat{\xi}$ can be negative, in which case we estimate ξ to be zero.

Suppose we infer that ξ is less than $\xi_0=.8$ if $\hat{\xi}<.8$, and if $\hat{\xi} \geq .8$, we decide $\xi \geq .8$. Let x_0 be the smallest integer such $\hat{\xi} \geq .8$. When $\xi=\xi_0-\delta^*$, the examinee's domain score is given by

$$\pi_1 = (1-\gamma)(\xi_0-\delta^*) + \varepsilon(1-\xi_0+\delta^*) \qquad [5.4]$$

and for $\xi = \xi_0 + \delta^*$, π is equal to

$$\pi_2 = (1-\gamma)(\xi_0+\delta^*) + \varepsilon(1-\xi_0-\delta^*). \qquad [5.5]$$

Thus, for $\xi = \xi_0 - \delta^*$

$$P(CD) = \sum_{x=0}^{x_0-1} \binom{n}{x} \pi_1^x (1-\pi_1)^{n-x} \qquad [5.6]$$

and for $\xi = \xi_0 + \delta^*$

$$P(CD) = \sum_{x=x_0}^{n} \binom{n}{x} \pi_2^x (1-\pi_2)^{n-x} \qquad [5.7]$$

To give some indication of the properties of using equation 5.3, we determined the smallest test length so that, simultaneously, equations 5.6 and 5.7 are at least $P^*=.9$ with $\delta^*=.1$. The results are reported in Table 1 for m=4,5 and various values of γ and ε.

In some cases, when the test length is formulated in terms of ξ, using equation 5.3 can substantially reduce the value of n over what would otherwise be required because we are, in essence, adjusting the passing score in a manner appropriate for what the values of γ and ε happen to be. This

result is to be expected. The important point made by the values of n in Table 1 is that the solution to the test length problem is highly sensitive to the values of γ and ε. Moreover, for the cases considered, the closer γ and ε are to zero, the smaller is the resulting value of n but it can be verified that this is not always the case. In practice, it is frequently assumed that $\gamma=0$ and that guessing is at random. It is generally conceded that this assumption is unrealistic, but often it is made anyway (e.g., Duncan, 1974). Weitzman (1970) proposes a procedure for ensuring guessing is at random. If this procedure is successfully implemented, we might substantially reduce the number of items that would otherwise be required.

6. Solutions Using Latent Structure Models

Since the test length is sensitive to the values of γ and ε, it would be helpful to have some method of estimating γ and ε or to have an estimate of ξ that does not assume guessing is at random. Under certain circumstances, such estimates are available (e.g., Anderson, 1954; Goodman, 1979; McHugh, 1956; Lazarsfeld and Henry, 1968). In this section, we consider test length when these methods are applied to estimate ξ.

Consistent with the previous section, it is assumed that an examinee is responding to a random sample of sets of equivalent items. Only situations involving pairs or triplets of equivalent items are considered. For a general approach to the case of items representing hierarchically related skills, see Dayton and Macready (1976). For an approach to determining the equivalency of item pairs the reader is referred to Baker and Hubert (1977) as well as

Hartke (1978). We also note that the usual method of judging the adequacy of a latent structure model is via the chi-squared goodness of fit test as illustrated by Macready and Dayton (1977).

Given that we are willing to make the assumptions necessary for the application of latent sturucture models, there are at least two technical problems when determining test length. The first is that we can no longer be certain that the P(CD) is minimized at either $\xi=\xi_0-\delta^*$ or $\xi=\xi_0+\delta^*$ unless, perhaps, we resort to an asymptotic argument or employ numerical techniques. The second is that there is no convenient method of determining, or even approximating, the smallest n so that equation 3.7 holds. No attempt is made to solve these problems; to be thorough it is necessary to indicate that these difficulties exist. In this section, we give brief consideration to whether latent structure models might be useful in reducing the number of items on the test that would otherwise be needed.

We begin by considering the case where two items per skill are used. For this situation, it is necessary to assume that one of the parameters γ or ε is known since, otherwise, the parameters are not uniquely determined and cannot be estimated. For present purposes, we assume that $\gamma=0$ when estimating ε.

As a comparison with the results on using the correction for guessing formula score, we used Monte Carlo methods to estimate the P(CD) using the values of n reported in Table 1 for the case m=4. The total number of skills on the test was set at n/2 or (n+1)/2, whichever gives an integral result. In each case we set $\xi_0=.8$ and made the estimates of the P(CD) with $\xi=\xi_0-\delta^*=.7$ and then with $\xi=\xi_0+\delta^*=.9$. The method used to estimate ξ is described by Wilcox (1979e). The results are reported in Table 2. As can be

seen, the latent structure model performs satisfactorily for $\gamma=0$ but even for γ slightly larger than zero the results do not support this approach when $\xi=.9$.

Next we considered using three items per skill for a total of 30 skills (and hence 90 items). In this case, the iterative procedure described by Goodman (1979) was used to approximate the maximum likelihood estimates of the parameters of the model. (It is no longer being assumed that $\gamma=0$.)

For the specific examinee being tested, let p_{ijk} be the probability of a particular pattern of responses on a randomly sampled triplet of equivalent items where a subscript of 0 or 1 corresponds to an incorrect and correct response respectively. For example, p_{011} denotes the probability of an incorrect on the first item and a correct on the other two. When applying Goodman's estimation, procedure one must estimate the p_{ijk}'s which are the cell probabilities of a multinomial distribution. Two estimation procedures were used. The first was the usual sample mean; the other was an estimate proposed by Fienberg and Holland (1973, see their equation 2.13).

It should be noted that when the sample mean is used to estimate the p_{ijk}'s, the solution to Goodman's equations (7), (8a),...,(8d) are maximum likelihood estimates of the parameters in the latent structure model. However, when the Fienberg-Holland estimate of the p_{ijk}'s are used, maximum likelihood estimates are no longer being obtained.

The results of our Monte Carlo studies are reported in Table 3. The columns headed MLE are the values of the P(CD) using Goodman's estimation procedure. The columns headed FH are the modified estimates based on the Fienberg-Holland estimate of the cell probabilities of a multinomial distribution. All indications are that the P(CD) is at least .9 when $\varepsilon=0$ or

.15 even for γ=.05 or .07. For all previous approaches, the P(CD) was considerably below .9 for γ=.05 and .07. However, for ε=.3 and particularly for ε=.4, the P(CD) is not very large. Note that the Fienberg-Holland estimate of the parameters in the multinomial distribution nearly always yields better results than those obtained with the sample mean estimate of the p_{ijk}'s.

No generalizations should be drawn from the Monte Carlo results reported in this section. Our goal was a more modest one: namely, to suggest what results might be obtained with latent structure models. The point is that it might be possible to take into account the errors γ and ε when comparing ξ to ξ_0 with a realistic, though perhaps large, number of items on the test. The results reported here are intended to motivate a more extensive investigation of the application of these models as well as the modified estimation procedure described above.

7. Solutions in Terms of ξ and a Population of Examinees

As was the case with percent correct true score, it is possible to formulate the test length problem in terms of ξ, the proportion of skills known by an examinee, where now we are concerned with the typical examinee among a population of examinees being tested. We note that when determining mastery of a single skill, rather than for a domain of skills, the solution described by Macready and Dayton (1977) might be applied.

For the special case ε>0 and γ=0 (or ε=0 and γ>0) the model proposed by Wilcox (1979e) might be used to obtain a point estimate of the probability of a false-positive or false-negative decision on the test. As before, if these

two errors are judged to be too high, one might increase the length of the test. Wilcox (1979e) illustrates how this might be done for the case of a single item per skill and so the details of the procedure are not discussed.

It is important to realize that in certain circumstances, the case of k>1 items per skill can be accomodated. Suppose, for example, that for each skill there are k items and that in each case the same decision rule (for instance all k items correct) is used to determine mastery. For a specific examinee and a sample of t skills (for a total of n=tk items), the probability of x mastery decisions is

$$\binom{t}{x} p^x (1-p)^{t-x} \qquad [7.1]$$

where p is the unknown probability of a mastery decision for a randomly sampled skill. Note that for this special case, $p=\xi+(1-\xi)\varepsilon_1\varepsilon_2\ldots\varepsilon_k$ where ε_i is the probability of guessing the ith item used to measure the skill. Let $\phi=(1-\xi)\varepsilon_1\ldots\varepsilon_k$ and assume that ξ and ϕ arise from a bivariate Dirichlet distribution. In other words, it is assumed that the joint probability density function of ξ and ϕ over the population of examinees is given by

$$\frac{\Gamma(v_1+v_2+v_3)}{\Gamma(v_1)\Gamma(v_2)\Gamma(v_3)} \xi^{v_1-1} \phi^{v_2-1} (1-\xi-\phi)^{v_3-1} \qquad [7.2]$$

where the v_i (i=1,2,3) are unknown parameters.

If t, the number of skills, is sufficiently large, as might be the case in a preliminary investigation, we can estimate ξ and the ε_i's for each examinee. (The effect of having a small number of skills appears to be unknown, cf. Wilcox, 1980). Once ξ and ϕ have been estimated for a random sample of examinees, we can estimate v_1, v_2 and v_3 in the manner described

by Wilcox (1979e). Substituting these estimates into the right-hand side of

$$f(x,\xi)=\binom{n}{x}B^{-1}(v_1,v_2,v_3)\sum_{w=0}^{x}\binom{x}{w}B(w+v_2,\ n-x+v_3)\xi^{x-w+v_1-1}(1-\xi)^{n-x+w+v_2+v_3-1} \qquad [7.3]$$

where $B(.,.,.)$ is the usual beta function, yields an estimate of the joint probability density function of x and ξ. If x_0 is the passing score of the test, the two possible errors are simply

$$\sum_{x=0}^{x_0-1}\int_{\xi_0}^{1} f(x,\xi)\,d\xi \qquad [7.4]$$

and

$$\sum_{x=x_0}^{n}\int_{0}^{\xi_0} f(x,\xi)\,d\xi \qquad [7.5]$$

which can be evaluated with subroutine BDTR in the IBM (1971) scientific subroutine package. Thus, we can determine the test length n=tk by adjusting t (and x_0) until equations 7.4 and 7.5 are sufficiently small.

Bounds on the Probability of an Error

If γ, the conditional probability that an examinee gives an incorrect response given that he/she knows the skill, is greater than zero and if ξ is estimated with a latent structure model, it is no longer clear how to estimate the probability of a false-positive and false-negative decision. However, if $\hat{\xi}$ is any estimate of ξ we can estimate the $P(\hat{\xi}<\xi_0)$ for a randomly selected examinee. This estimate is simply the proportion of examinees for whom $\hat{\xi}<\xi_0$. Moreover, if $\hat{\xi}$ is consistent, we can estimate the mean and variance

of ξ over the population of examinees (Wilcox, 1979e). Thus, following Wilcox (1979c) bounds on the two error types can be estimated. As previously explained, these bounds give us information about whether there are enough items on the test.

8. Solutions Using Latent Trait Models

The third general approach to the test length problem is based on a latent trait model. For a specific examinee the probability of a correct response to a test item, say $p_\tau(\zeta)$, is viewed as a function of ζ, $-\infty < \zeta < \infty$, the examinee's "ability" level, and the vector τ which consists of parameters that characterize the item. Lord (1974) interprets $p_\tau(\zeta)$ as a relative frequency over randomly selected test questions all having the same vector of τ values.

Several forms for $p_\tau(\zeta)$ have been proposed (see, e.g., Hambleton and Cook, 1977). For a recent review of latent trait models, the reader is referred to Hambleton, et al., (1978). Familiarity with this review is assumed henceforth.

Birnbaum (1968) considers the classification of examinees in some detail. As was the case with the two types of true score previously considered, it is assumed that two ability levels have been specified, say ζ_1 and ζ_2 with the idea that if $\zeta \leq \zeta_1$ or if $\zeta \geq \zeta_2$, we want to be reasonably certain of making a correct decision about whether the examinee's true score is large or small.

Rather than formulate the test length problem in terms of the probability of a correct decision, Birnbaum chooses n so that the probabilities associated

with the two possible errors do not exceed prespecified values. That is, we choose n so that simultaneously

$$P(x \geq x_0 \mid \zeta = \zeta_1) \leq \alpha^* \qquad [8.1]$$

and

$$P(x < x_0 \mid \zeta = \zeta_2) \leq \beta^* \qquad [8.2]$$

where x_0 is the passing score and α^* and β^* are preassigned constants. Note that x need not be a number correct score. Birnbaum (1968, eq. 19.5.13) derives an approximation to the minimal n satisfying equations 8.1 and 8.2 given by

$$n^{\frac{1}{2}} = \frac{\Phi^{-1}(1-\alpha^*)[p_\beta(\zeta_1)(1-p_\beta(\zeta_1))]^{\frac{1}{2}} - \Phi^{-1}(\beta^*)[p_\beta(\zeta_2)(1-p_\beta(\zeta_2))]^{\frac{1}{2}}}{p_\beta(\zeta_2) - p_\beta(\zeta_1)} \qquad [8.3]$$

where Φ^{-1} is the inverse function of the standard normal cumulative distribution. (Actually this expression for n differs slightly from Birnbaum's which apparently has a typographical error.) To apply this solution one must already have an estimate of the function $p_\tau(\zeta)$. The solution also makes the highly restrictive assumption of equivalent items, i.e., every item has the same values for τ.

As with all the probability models in this paper that attempt to make inferences about what an examinee knows beyond the observed responses on a test, there are several technical issues that remain to be resolved--not the least of which is a guideline on when one form of $p_\tau(\zeta)$ is to be preferred over another. Some of these issues are discussed by Hambleton et al. (1978).

Certainly latent trait models deserve careful study and consideration. When measuring achievement, there are at least two issues that deserve a

special comment. The first, which is raised by Baker (1977), is whether latent trait models are even appropriate at all. As Baker puts it, latent trait theory is the culmination of the work on the measurement of ability, begun by Binet, that was the major focus of psychometrics in the 1920's, 30's and 40's. He goes on to point out that the educational problems of an earlier era are not the problems of the 1970's and 80's. The major trend in educational measurement today is one of instructionally related testing. Moreover, the problems arising from the individualization of instruction are very different from those of ability measurement.

A more specific problem with latent trait models that needs to be considered, is what we do with the items that do not fit the model. From comments made by Gustafsson (1979), it would seem that many such items might exist when the Rasch model is assumed to hold. If these items really do represent a skill associated with a particular instructional program, it may be of interest to determine whether an examinee has mastered the skill even if the item does not fit a particular latent trait model. Ron Hambleton pointed out that this problem should be addressed at the test development stage; if we have evidence that the items measure the objectives, and if the model does not fit, we should throw out the model--not the items.

In terms of the present paper, the question is: if we chhose not to ignore the items that do not fit a latent trait model, what do we do with these items and how do we relate our actions to the problem of test length? It should be mentioned, however, that under certain circumstances an argument has been made in favor of latent trait models over item sampling models (Wood, 1976). In addition Messick, (1975, p. 957) argues

that all measurement should be construct referenced and that a measure estimate how much of a trait an individual possesses. Nevertheless, the issues of what to do, if anything, with items that do not fit a latent trait model has yet to be discussed.

A Concluding Remark

Perhaps the most important point of this paper is that there is no magic number-or even magic formula-for determining test length. Even within the seemingly narrow problem of comparing an examinee's true score to some constant, there are many approaches to the problem. Moreover, in terms of which true score to use, it is not at all clear as to what extent the three types considered here are in competition with one another. For the moment, the best we can do is to be very precise about what we want to determine, consider what assumptions we are willing to make, and act accordingly.

REFERENCES

Aitchison, J, & Dunsmore, I. R. Statistical Prediction Analysis. Cambridge University Press, 1975.

Andrews, D. F.; Bickel, P. J., Hampel, F. R., Huber, P. J., Rogers, W. H., & Tukey, J. W. Robust estimates of location. Princeton: Princeton University Press, 1972.

Baker, F. B. Advances in item analysis. Review of Educational Research, 1977, 47, 151-178.

Baker, F. B., & Hubert, L. J. Inference procedures for ordering theory. Journal of Educational Statistics, 1977, 2, 217-233.

Bechhofer, R. E. A single-sample multiple decision procedure for ranking means of normal populations with known variances. Annals of Mathematical Statistics, 1954, 25, 16-39.

Bergan, J. R., Camelli, A. A., & Luiten, J. W. Mastery assessment with latent class and quasi-independence models representing homogeneous item domains. Journal of Educational Statistics, 1980, 5, 65-81.

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord, & M. R. Novick (Eds.) Statistical Theories of Mental Test Scores. Reading, Mass.: Addison-Wesley, 1968.

Blischke, W. R. Estimating the parameters of mixtures of binomial distributions. Journal of the American Statistical Association, 1964, 59, 510-528.

Cochran, W. G. Some methods for strengthening the common χ^2 tests. Biometrica, 1954, 10, 417-451.

Dayton, C. M., & Macready, G. B. A probabilistic model for validation of behavioral hierarchies. Psychometrika, 1976, 41, 189-204.

Dempster, A. P. New approaches for reasoning towards posterior distributions based on sample data. Annals of Mathematical Statistics, 1966, 37, 355-374.

Dempster, A. P. Upper and lower probabilities induced by a multivalued mapping. Annals of Mathematical Statistics, 1967, 36, 325-339.

Dempster, A. P. Upper and lower probabilities generated by a random closed interval. Annals of Mathematical Statistics, 1968, 39, 957-966(a).

Dempster, A. P. A generalization of Bayesian inference. Journal of the Royal Statistical Society, Ser. B, 1968, 30, 205-232.

Duncan, G. T. An empirical Bayes approach to scoring multiple-choice tests in the misinformation model. Journal of the American Statistical Association, 1974, 69, 50-57.

Fhanér, S. Item sampling and decision making in achievement testing. British Journal of Mathematical and Statistical Psychology, 1974, 27, 172-175.

Fienberg, S. E., & Holland, P. W. Simultaneous estimation of multinomial cell probabilities. Journal of the American Statistical Association, 1973, 68, 683-691.

Freeman, M. F., & Tukey, J. W. Transformations related to the angular and the square root. The Annals of Mathematical Statistics, 1950, 21, 607-611.

Gelfand, A., & Thomas, D. Discrimination between the binomial and hypergeometric models. Communications in Statistics A, Theory and Methods, 1976, 18, 225-240.

Godambe, V. P., & Sprott, D. A. Foundations of Statistical Inference. Toronto: Holt, Rinehart and Winston, 1971.

Goodman, L. A. Exploratory latent structure analysis using both indentifiable and unidentifiable models. Biometrika, 1974, 61, 215-231.

Goodman, L. A. On the estimation of parameters in latent structure analysis. Psychometrika, 1979, 44, 123-128.

Goodman, L. A., & Kruskal, W. H. Measures of association for cross classifications II: Further discussion and references. Journal of the American Statistical Association, 1959, 54, 123-163.

Gustafsson, J. Testing and obtaining fit of data to the Rasch model. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, 1979.

Hambleton, R., & Cook, L. Latent trait models and their use in the analysis of educational test data. Journal of Educational Measurement, 1977, 14, 75-96.

Hambleton, R. K., Swaminathan, H., Algina, J., & Coulson, D. B. Criterion-referenced testing and measurement: A review of technical issues and developments. Review of Educational Research, 1978, 48, 1-47.

Hambleton, R. K., Swaminathan, H., Cook, L. L., Eignor, D. R., & Gifford, J. A. Developments in latent trait theory: Models, technical issues, and applications. Review of Educational Research, 1978, 48, 467-510.

Hartke, A. R. The use of latent partition analysis to identify homogeneity of an item population. Journal of Educational Measurement, 1978, 15, 43-47.

Harris, C. W., & Pearlman, A. P. An index for a domain of completion or short answer items. Journal of Educational Statistics, 1978, 3, 285-304.

Huang, W. Bayes approach to a problem in partitioning k normal populations. Bulletin of the Institute of Mathematics Academia Sinica, 1975, 3, 87-97.

Huynh, H. Statistical inference for false positive and false negative error rates in mastery. Psychometrika, 1980, 45, 107-120.

Huynh, H. Statistical consideration of mastery scores. Psychometrika, 1976, 41, 65-78.

IBM Application Program, System/360. Scientific subroutines package (360-CM-03X) Version III, programmer's manual. White Plains, New York: IBM Corporation Technical Publications Department, 1971.

Johnson, N. L., & Kotz, S. Urn models and their application. New York: Wiley, 1977.

Katz, L. Unified treatment of a broad class of discrete probability distributions. In G. P. Patil (Ed.) Classical and Contagious Discrete Distributions. New York: Pergamon Press, 1963.

Keats, J. A., & Lord, F. M. A theoretical distribution for mental test scores. Psychometrika, 1962, 27, 59-72.

Kendall, M. G., & Stuart, A. The advanced theory of statistics, Vol. 2, New York: Hafner, 1973.

Lam, K., & Chiu, W. K. On the probability of correctly selecting the best of several normal populations. Biometrika, 1976, 63, 410-411.

Lazarsfeld, P. F., & Henry, N. W. Latent structure analysis. New York: Houghton Mifflin, 1968.

Livingston, S. A., & Wingersky, M. S. Assessing the reliability of tests used to make pass/fail decisions. Journal of Educational Measurement, 1979, 16, 247-260.

Lord, F. M. A strong true-score theory, with applications. Psychometrika, 1965, 30, 239-270.

Lord, F. M. Estimating true-score distributions in psychological testing (an empirical Bayes estimation problem). Psychometrika, 1969, 34, 259-299.

Lord, F. M., & Novick, M. R. Statistical theories of mental test scores. Reading, Mass.: Addison - Wesley, 1968.

Lord, F. M. Individualized testing and item characteristic curve theory. In D. H. Krantz, R. C. Atkinson, R. D. Lace, & P. Suppes (Eds.) Contemporary developments in mathematical psychology, Vol. II. San Francisco, CA: Freeman, 1974.

Lumsden, J. Test Theory, Annual Review of Psychology, 1976, 27, 251-280.

Maritz, J. S. Empirical Bayes methods. London: Methuen, 1970.

Macready, G. B., & Dayton, C. M. The use of probabilistic models in the assessment of mastery. Journal of Educational Statistics, 1977, 2, 99-120.

Marks, E., & Noll, G. A. Procedures and criteria for evaluating reading and listening comprehension tests. Educational and Psychological Measurement, 1967, 27, 335-348.

McHugh, R. B. Efficient estimation and local identification in latent class analysis. Psychometrika, 1956, 21, 331-347.

Messick, S. The standard problem: Meaning and values in measurement and evaluation. American Psychologist, 1975, 30, 955-966.

Millman, J. Passing scores and test lengths for domain-referenced measures. Review of Educational Research, 1973, 43, 205-216.

Molemnaar, W. On Bayesian formula scores for random guessing in multiple choice tests. British Journal of Mathematical and Statistical Psychology, 1977, 30, 79-89.

Morgan, G. A criterion-referenced measurement model with corrections for guessing and carelessness. The Australian Council for Educational Research Limited, Occasional paper no. 13, Jenkin Buxton Printers. Ptg Ltd, 1979.

Murray, G. D. A note on the estimation of probability density fundtions. Biometrika, 1977, 64, 150-151.

Novick, M. R. High school attainment: An example of a computer-assisted Bayesian approach to data analysis. International Statistical Review, 1973, 41, 264-271.

Novick, M. R., & Jackson, P. H. Statistical Methods for Educational and Psychological Research. New York: McGraw-Hill, 1974.

Novick, M. R., & Lewis, C. Prescribing test length for criterion-referenced measurement. In C. W. Harris, M. C. Alkin, & W. J. Popham (Eds.), Problems in criterion-referenced measurement. CSE Monograph series in Evaluation, No. 3. Los Angeles: Center for the Study of Evaluation, University of California, 1974.

Novick, M. R., Lewis, C., & Jackson, P. H. The estimation of proportions in m groups. Psychometrika, 1973, 38, 19-46.

Reulecke, W. A. A statistical analysis of deterministic theories. In H. Spada & F. Kempf (Eds.) Structural Models of Thinking and Learning. Bern: Huber, 1977.

Rustagi, J. S. Variational methods in statistics. New York: Academic Press, 1976.

Skibinsky, M. Sharp upper bounds for probability on an interval when the first three moments are known. The Annals of Statistics, 1976, 4, 187-213.

Skibinsky, M. The maximum probability on an interval when the mean and variance are known. Sankhya, 1977, Series A, 39, 144-159.

Springer, M. D. The algebra of random variables. New York: Wiley, 1979.

Tarone, R. E. Testing the goodness of bit of the binomial distribution. Biometrika, 1979, 66, 585-590.

Tong, Y. L., & Wetzell, D. E. On the behaviour of the probability function for selecting the best normal population. Biometrika, 1979, 66, 174-176.

van den Brink, W. P., & Koele, P. Item sampling, guessing and decision-making in achievement testing. British Journal of Mathematical and Statistical Psychology, 1980, 33, 104-108.

Van der Linden, W. Forgetting, guessing, and mastery: The Macready and Dayton models revisited and compared with a latent trait approach. Journal of Educational Statistics, 1978, 3, 305-317.

Von Mises, R. A mathematical theory of probability and statistics. New York: Academic Press, 1964.

Weitzman, R. A. Ideal multiple-choice items. Journal of the American Statistical Association, 1970, 65, 71-89.

Wilcox, R. R. Some results and comments on using latent structure models to measure achievement. Educational and Psychological Measurement, 1980, in press.

Wilcox, R. R. Applying ranking and selection techniques to determine the length of a mastery test. Educational and Psychological Measurement, 1979, 39, 13-22(a).

Wilcox, R. R. Comparing examinees to a control. Psychometrika, 1979, 44, 55-68(b).

Wilcox, R. R. On false-positive and false-negative decisions with a mastery test. Journal of Educational Statistics, 1979, 4, 59-73(c).

Wilcox, R. R. An approach to measuring the achievement or proficiency of an examinee. Applied Psychological Measurement, 1980, in press.

Wilcox, R. R. Achievement tests and latent structure models. British Journal of Mathematical and Statistical Psychology, 1979, 32, 61-71(e).

Wilcox, R. R. Estimating the parameters of the beta-binomial distribution. Education and Psychological Measurement, 1979, 39, 527-535(f).

Wilcox, R. R. Estimating the likelihood of a false-positive or false-negative decision with a mastery test: An empirical Bayes approach. Journal of Educational Statistics, 1977, 2, 289-307.

Wood, R. Trait measurement and item banks. In D. de Gruijter & L. van der Kamp (Eds.) Advances in Psychological and Educational Measurement. New York: Wiley, 1976.